Towards a new integrated uniform production system for business statistics at Statistics Netherlands

Automatic data editing with multiple data sources

Sander Scholtus, Wilco de Jong, Anita Vaasen-Otten & Frank Aelen

3-10-2022

Automatic data editing in a new integrated uniform production system

- Goal is an efficient and flexible integrated uniform production system for business statistics
- Multiple data sources combined in one process
- Automatic data editing important part of the new process to retrieve efficiency
- Challenges:
  - Which data source is the best?
  - How to resolve inconsistencies between data sources?
  - Is it possible to edit all data sources at the same time?
  - Is automatic data editing more efficient when using multiple data sources?

***Manual top-down data editing is described in a companion paper**

Automatic editing: review

- Three approaches for automatic editing
  - Deductive correction (IF-THEN rules or simple algorithms)
    - Explicit correction rules
      (e.g., IF Turnover_t / Turnover_{t-1} > 300 THEN Turnover_t := Turnover_t / 1000)
  - Error localization based on the Fellegi-Holt paradigm
    - Uses edit rules (e.g., Turnover ≥ 0; Total = Sum of subtotals)
    - Minimizes the (weighted) number of edited values for each unit
    - Mathematical formulation: mixed-integer minimization problem
  - Error localization with general edit operations
    - Uses edit rules
    - Minimizes the (weighted) number of edit operations for each unit
    - An edit operation may affect more than one value, in a pre-specified way
    - Can be seen as a generalization of the Fellegi-Holt paradigm

Automatic editing: review

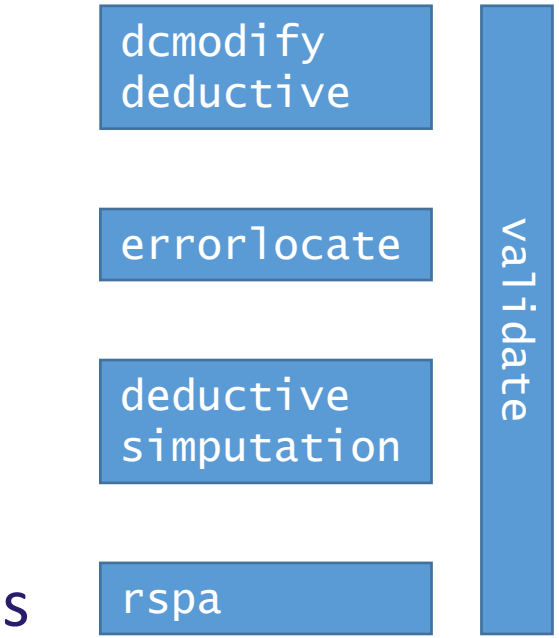
- Potential use of edit operations for automatic editing across statistics
  - Example: using auxiliary register data during editing of SBS data

	raw data
SWL: Total wage costs	*8 000*
SBS: Gross wages	*10 000*
SBS: Social benefits	2 000
SBS: Pension costs	1 000
SBS: Other personnel costs	1 000
SBS: Total personnel costs	14 000

Automatic editing: review

- Typical automatic editing process:

1. Deductive correction (systematic errors)
2. Error localization (other errors)
3. Imputation of missing values
4. Adjustment of imputed values to satisfy all edit rules



Automatic editing across data sources

Source 1

observed variables

$$\mathbf{x}_i^{(1)} = \left(x_{i1}^{(1)}, \ldots, x_{iJ_1}^{(1)}\right)'$$

edit rules

$$\mathbf{A}_u^{(1)}\mathbf{x}_i^{(1)} \leq \mathbf{b}_u^{(1)}$$

$$\mathbf{A}_e^{(1)}\mathbf{x}_i^{(1)} = \mathbf{b}_e^{(1)}$$

common variables

$$\mathbf{y}_i^{(1)} = \left(y_{i1}^{(1)}, \ldots, y_{iL_1}^{(1)}\right)'$$

derivations / edit rules

$$y_{il}^{(1)} = \sum_j c_{lj}^{(1)} x_{ij}^{(1)}$$

Source 2

observed variables

$$\mathbf{x}_i^{(2)} = \left(x_{i1}^{(2)}, \ldots, x_{iJ_2}^{(2)}\right)'$$

edit rules

$$\mathbf{A}_u^{(2)}\mathbf{x}_i^{(2)} \leq \mathbf{b}_u^{(2)}$$

$$\mathbf{A}_e^{(2)}\mathbf{x}_i^{(2)} = \mathbf{b}_e^{(2)}$$

common variables

$$\mathbf{y}_i^{(2)} = \left(y_{i1}^{(2)}, \ldots, y_{iL_2}^{(2)}\right)'$$

derivations / edit rules

$$y_{il}^{(2)} = \sum_j c_{lj}^{(2)} x_{ij}^{(2)}$$

...

...

Source P

observed variables

$$\mathbf{x}_i^{(P)} = \left(x_{i1}^{(P)}, \ldots, x_{iJ_P}^{(P)}\right)'$$

edit rules

$$\mathbf{A}_u^{(P)}\mathbf{x}_i^{(P)} \leq \mathbf{b}_u^{(P)}$$

$$\mathbf{A}_e^{(P)}\mathbf{x}_i^{(P)} = \mathbf{b}_e^{(P)}$$

common variables

$$\mathbf{y}_i^{(P)} = \left(y_{i1}^{(P)}, \ldots, y_{iL_P}^{(P)}\right)'$$

derivations / edit rules

$$y_{il}^{(P)} = \sum_j c_{lj}^{(P)} x_{ij}^{(P)}$$

'true' values

$$\mathbf{z}_i = (z_{i1}, \ldots, z_{iL})'$$

edit rules

$$\left|y_{il}^{(p)} - z_{il}\right| \leq \epsilon_l^* |z_{il}|$$

$$\mathbf{A}_u\mathbf{z}_i \leq \mathbf{b}_u$$

$$\mathbf{A}_e\mathbf{z}_i = \mathbf{b}_e$$

Task: adjust the data $\left(\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(P)}, \mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}, \mathbf{z}_i\right)$ as necessary so that all edit rules are satisfied

Example

across data sources:

variable	description	data source	orig. value
$y_{i1}^{(1)}$	Total turnover, definition 1	SBS	16 091
$y_{i1}^{(2)}$		ProdCom	2 504
z_{i1}		'true' value	.
$y_{i2}^{(1)}$	Total turnover, definition 2	SBS	16 091
$y_{i2}^{(3)}$		SFKO	18 496
$y_{i2}^{(4)}$		STS-admin	19 386
$y_{i2}^{(5)}$		STS-survey	16 610
z_{i2}		'true' value	.
$y_{i3}^{(1)}$	Total industrial production	SBS	16 091
$y_{i3}^{(2)}$		ProdCom	15 582
$y_{i3}^{(5)}$		STS-survey	16 610
z_{i3}		'true' value	.
$y_{i4}^{(1)}$	Domestic industrial production	SBS	16 091
$y_{i4}^{(2)}$		ProdCom	15 582
z_{i4}		'true' value	.
$y_{i5}^{(1)}$	Domestic industrial production of goods	SBS	16 091
$y_{i5}^{(2)}$		ProdCom	2 504
z_{i5}		'true' value	.

Structural Business Statistics (SBS):

variable	description	orig. value
$x_{i1}^{(1)}$	Total turnover (SBS)	16 091
$x_{i2}^{(1)}$	Excises paid	0
$x_{i3}^{(1)}$	On-charged freight costs	0
$x_{i4}^{(1)}$	Total industrial production of goods (SBS)	16 091
$x_{i5}^{(1)}$	Domestic industrial production of goods (SBS)	16 091
$x_{i6}^{(1)}$	Total industrial production of services (SBS)	0
$x_{i7}^{(1)}$	Domestic industrial production of services (SBS)	0
	… over 100 other variables …	

Derivation of common variables in SBS:

$$y_{i1}^{(1)} = x_{i1}^{(1)} - x_{i2}^{(1)} - x_{i3}^{(1)}$$
$$y_{i2}^{(1)} = x_{i1}^{(1)}$$
$$y_{i3}^{(1)} = x_{i4}^{(1)} + x_{i6}^{(1)} - x_{i2}^{(1)} - x_{i3}^{(1)}$$
$$y_{i4}^{(1)} = x_{i5}^{(1)} + x_{i7}^{(1)} - x_{i2}^{(1)} - x_{i3}^{(1)}$$
$$y_{i5}^{(1)} = x_{i5}^{(1)} - x_{i2}^{(1)} - x_{i3}^{(1)}$$

Automatic editing across data sources

- Proposal: use a three-step approach
  1. Automatic editing of common variables across data sources
     - Identify errors in $\left(\mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}, \mathbf{z}_i\right)$ using edit rules across data sources
  2. Fix 'true' values and derive additional edit rules on common variables:
     - Impute 'true' values $\mathbf{z}_i$ consistently with edit rules
     - Derive additional edit rules for $\left(\mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}\right)$ from $\left|y_{il}^{(p)} - z_{il}\right| \leq \epsilon_l^* |z_{il}|$
  3. Automatic editing within each individual data source
     - Identify errors and impute new values in $\left(\mathbf{x}_i^{(p)}, \mathbf{y}_i^{(p)}\right)$ so all edit rules are satisfied
- In Step 1 and Step 3 both deductive correction and error localization can be used

Example: Step 1

Step 1: Automatic editing of common variables across data sources

variable	description	data source	weight	orig. value
$y_{i1}^{(1)}$	Total turnover, definition 1	SBS	5	16 091
$y_{i1}^{(2)}$		ProdCom	3	2 504
z_{i1}		'true' value	.	.
$y_{i2}^{(1)}$	Total turnover, definition 2	SBS	5	16 091
$y_{i2}^{(3)}$		SFKO	3	18 496
$y_{i2}^{(4)}$		STS-admin	7	19 386
$y_{i2}^{(5)}$		STS-survey	3	16 610
z_{i2}		'true' value	.	.
$y_{i3}^{(1)}$	Total industrial production	SBS	4	16 091
$y_{i3}^{(2)}$		ProdCom	6	15 582
$y_{i3}^{(5)}$		STS-survey	3	16 610
z_{i3}		'true' value	.	.
$y_{i4}^{(1)}$	Domestic industrial production	SBS	4	16 091
$y_{i4}^{(2)}$		ProdCom	8	15 582
z_{i4}		'true' value	.	.
$y_{i5}^{(1)}$	Domestic industrial production of goods	SBS	4	16 091
$y_{i5}^{(2)}$		ProdCom	8	2 504
z_{i5}		'true' value	.	.

Edit rules for 'true' values:

$$z_{i2} \geq z_{i1}$$

$$z_{i1} \geq z_{i3}$$

$$z_{i3} \geq z_{i4}$$

$$z_{i4} \geq z_{i5}$$

$$z_{i5} \geq 0$$

Edit rules on common variables:

$$\left| y_{il}^{(p)} - z_{il} \right| \leq 0.05 \times |z_{il}|$$

Example: Step 2

Step 2: Fix ‘true’ values and derive additional edit rules on common variables

variable	description	data source	weight	orig. value	step 1
$y_{i1}^{(1)}$	Total turnover, definition 1	SBS	5	16 091	16 091
$y_{i1}^{(2)}$		ProdCom	3	2 504	.
z_{i1}		‘true’ value	.	.	.
$y_{i2}^{(1)}$	Total turnover, definition 2	SBS	5	16 091	.
$y_{i2}^{(3)}$		SFKO	3	18 496	18 496
$y_{i2}^{(4)}$		STS-admin	7	19 386	19 386
$y_{i2}^{(5)}$		STS-survey	3	16 610	.
z_{i2}		‘true’ value	.	.	.
$y_{i3}^{(1)}$	Total industrial production	SBS	4	16 091	16 091
$y_{i3}^{(2)}$		ProdCom	6	15 582	15 582
$y_{i3}^{(5)}$		STS-survey	3	16 610	16 610
z_{i3}		‘true’ value	.	.	.
$y_{i4}^{(1)}$	Domestic industrial production	SBS	4	16 091	16 091
$y_{i4}^{(2)}$		ProdCom	8	15 582	15 582
z_{i4}		‘true’ value	.	.	.
$y_{i5}^{(1)}$	Domestic industrial production of goods	SBS	4	16 091	.
$y_{i5}^{(2)}$		ProdCom	8	2 504	2 504
z_{i5}		‘true’ value	.	.	.

Edit rules on common variables:

$$\left|y_{il}^{(p)} - z_{il}\right| \leq 0.05 \times |z_{il}|$$

Find feasible intervals for z_{il}:

$$15\,819 \leq z_{i1} \leq 16\,938$$

$$18\,463 \leq z_{i2} \leq 19\,469$$

$$15\,819 \leq z_{i3} \leq 16\,402$$

$$15\,325 \leq z_{i4} \leq 16\,402$$

$$2\,385 \leq z_{i5} \leq 2\,636$$

Example: Step 2

Step 2: Fix ‘true’ values and derive additional edit rules on common variables

variable	description	data source	weight	orig. value	step 1	step 2
$y_{i1}^{(1)}$	Total turnover, definition 1	SBS	5	16 091	16 091	16 091
$y_{i1}^{(2)}$		ProdCom	3	2 504	.	.
z_{i1}		‘true’ value	.	.	.	**16 091**
$y_{i2}^{(1)}$	Total turnover, definition 2	SBS	5	16 091	.	.
$y_{i2}^{(3)}$		SFKO	3	18 496	18 496	18 496
$y_{i2}^{(4)}$		STS-admin	7	19 386	19 386	19 386
$y_{i2}^{(5)}$		STS-survey	3	16 610	.	.
z_{i2}		‘true’ value	.	.	.	**19 386**
$y_{i3}^{(1)}$	Total industrial production	SBS	4	16 091	16 091	16 091
$y_{i3}^{(2)}$		ProdCom	6	15 582	15 582	15 582
$y_{i3}^{(5)}$		STS-survey	3	16 610	16 610	16 610
z_{i3}		‘true’ value	.	.	.	**16 091**
$y_{i4}^{(1)}$	Domestic industrial production	SBS	4	16 091	16 091	16 091
$y_{i4}^{(2)}$		ProdCom	8	15 582	15 582	15 582
z_{i4}		‘true’ value	.	.	.	**15 582**
$y_{i5}^{(1)}$	Domestic industrial production of goods	SBS	4	16 091	.	.
$y_{i5}^{(2)}$		ProdCom	8	2 504	2 504	2 504
z_{i5}		‘true’ value	.	.	.	**2 504**

Edit rules on common variables:

$$\left|y_{il}^{(p)} - z_{il}\right| \leq 0.05 \times |z_{il}|$$

Feasible intervals for z_{il}:

$15\,819 \leq z_{i1}^{(1)} \leq 16\,938$
$15\,286 \leq y_{i1}^{(1)} \leq 16\,896$
$18\,463 \leq z_{i2}^{(1)} \leq 19\,469$
$18\,417 \leq y_{i2}^{(1)} \leq 20\,355$
$15\,819 \leq z_{i3} \leq 16\,402$
$15\,286 \leq y_{i3}^{(1)} \leq 16\,896$
$15\,325 \leq z_{i4} \leq 16\,402$
$14\,803 \leq y_{i4}^{(1)} \leq 16\,361$
$2\,385 \leq z_{i5} \leq 2\,636$
$2\,379 \leq y_{i5}^{(1)} \leq 2\,629$

Example: Step 3

Step 3: Automatic editing within an individual data source (SBS)

variable	description	weight	orig. value	step 2
$x_{i1}^{(1)}$	Total turnover (SBS)	5	16 091	16 091
$x_{i2}^{(1)}$	Excises paid	3	0	0
$x_{i3}^{(1)}$	On-charged freight costs	4	0	0
$x_{i4}^{(1)}$	Total industrial production of goods (SBS)	1	16 091	16 091
$x_{i5}^{(1)}$	Domestic industrial production of goods (SBS)	1	16 091	16 091
$x_{i6}^{(1)}$	Total industrial production of services (SBS)	1	0	0
$x_{i7}^{(1)}$	Domestic industrial production of services (SBS)	2	0	0
	… over 100 other variables …			

variable	description	weight	orig. value	step 2
$y_{i1}^{(1)}$	Total turnover, definition 1 (SBS)	0.01	16 091	16 091
$y_{i2}^{(1)}$	Total turnover, definition 2 (SBS)	0.01	16 091	.
$y_{i3}^{(1)}$	Total industrial production (SBS)	0.01	16 091	16 091
$y_{i4}^{(1)}$	Domestic industrial production (SBS)	0.01	16 091	16 091
$y_{i5}^{(1)}$	Domestic industrial production of goods (SBS)	0.01	16 091	.

Edit rules on common variables for SBS:

$$15\,286 \leq y_{i1}^{(1)} \leq 16\,896$$

$$18\,417 \leq y_{i2}^{(1)} \leq 20\,355$$

$$15\,286 \leq y_{i3}^{(1)} \leq 16\,896$$

$$14\,803 \leq y_{i4}^{(1)} \leq 16\,361$$

$$2\,379 \leq y_{i5}^{(1)} \leq 2\,629$$

In addition:

- Edit rules relating internal SBS variables to common variables
- Internal SBS edit rules

Automatic editing across data sources

- During step 3, additional errors in common variables $\mathbf{y}_i^{(p)}$ may be identified using internal edit rules of data source p
  - These can be resolved automatically within data source p provided that no new inconsistencies are introduced with respect to $\left|y_{il}^{(p)} - z_{il}\right| \leq \epsilon_l^* |z_{il}|$
- Use score function to identify units with large inconsistencies between $\left(\mathbf{x}_i^{(p)}, \mathbf{y}_i^{(p)}\right)$ and other sources
  - For those units, automatic editing across data sources may not give good results

Concluding remarks

- Three-step approach tested in a Proof of Concept
- Main findings:
  - Technically feasible within current IT environment (unlike one-step approach)
  - Use of edit operations leads to greater flexibility and enables automatic editing to mimic commonly-used data editing strategies of subject-matter experts
  - Quality of automatically edited data is still quite low. To improve this:
    - Identify and explain possible systematic differences between data sources
    - Reach consensus among experts about 'ideal' way to resolve inconsistencies
    - Ongoing process to learn from experiences during manual editing
- Near future: Pilot for interactive and automatic editing across data sources